PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 7 :
C12N 15/12, C07K 14/47, C12N 15/63, A61K 38/17, C07K 16/18, A61K 39/395, C12Q 1/68, A61K 48/00

A2

(11) International Publication Number: WO 00/50588

(43) International Publication Date: 31 August 2000 (31.08.00)

(21) International Application Number: PCT/US00/02595

(22) International Filing Date: 1 February 2000 (01.02.00)

(30) Priority Data:
09/255,381 22 February 1999 (22.02.99) US

(71) Applicant *(for all designated States except US)*: INCYTE PHARMACEUTICALS, INC. [US/US]; 3174 Porter Drive, Palo Alto, CA 94304 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: WALKER, Michael, G. [CA/US]; Unit 80, 1050 Borregas Avenue, Sunnyvale, CA 94089 (US). VOLKMUTH, Wayne [US/US]; 783 Roble Avenue #1, Menlo Park, CA 94025 (US). KLINGLER, Tod, M. [US/US]; 28 Dover Court, San Carlos, CA 94070 (US). LAL, Preeti [IN/US]; 2382 Lass Drive, Santa Clara, CA 95054 (US).

(74) Agents: MURRY, Lynn, E. et al.; Incyte Pharmaceuticals, Inc., 3174 Porter Drive, Palo Alto, CA 94304 (US).

(81) Designated States: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published
Without international search report and to be republished upon receipt of that report.

(54) Title: GENES ASSOCIATED WITH DISEASES OF THE COLON

(57) Abstract

The invention provides colon cancer genes and polypeptides encoded by those genes. The invention also provides expression vectors, host cells, and antibodies. The invention also provides methods for diagnosing, treating or preventing diseases of the colon.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania
AM	Armenia
AT	Austria
AU	Australia
AZ	Azerbaijan
BA	Bosnia and Herzegovina
BB	Barbados
BE	Belgium
BF	Burkina Faso
BG	Bulgaria
BJ	Benin
BR	Brazil
BY	Belarus
CA	Canada
CF	Central African Republic
CG	Congo
CH	Switzerland
CI	Côte d'Ivoire
CM	Cameroon
CN	China
CU	Cuba
CZ	Czech Republic
DE	Germany
DK	Denmark
EE	Estonia
ES	Spain
FI	Finland
FR	France
GA	Gabon
GB	United Kingdom
GE	Georgia
GH	Ghana
GN	Guinea
GR	Greece
HU	Hungary
IE	Ireland
IL	Israel
IS	Iceland
IT	Italy
JP	Japan
KE	Kenya
KG	Kyrgyzstan
KP	Democratic People's Republic of Korea
KR	Republic of Korea
KZ	Kazakstan
LC	Saint Lucia
LI	Liechtenstein
LK	Sri Lanka
LR	Liberia
LS	Lesotho
LT	Lithuania
LU	Luxembourg
LV	Latvia
MC	Monaco
MD	Republic of Moldova
MG	Madagascar
MK	The former Yugoslav Republic of Macedonia
ML	Mali
MN	Mongolia
MR	Mauritania
MW	Malawi
MX	Mexico
NE	Niger
NL	Netherlands
NO	Norway
NZ	New Zealand
PL	Poland
PT	Portugal
RO	Romania
RU	Russian Federation
SD	Sudan
SE	Sweden
SG	Singapore
SI	Slovenia
SK	Slovakia
SN	Senegal
SZ	Swaziland
TD	Chad
TG	Togo
TJ	Tajikistan
TM	Turkmenistan
TR	Turkey
TT	Trinidad and Tobago
UA	Ukraine
UG	Uganda
US	United States of America
UZ	Uzbekistan
VN	Viet Nam
YU	Yugoslavia
ZW	Zimbabwe

GENES ASSOCIATED WITH DISEASES OF THE COLON

TECHNICAL FIELD

The invention relates to seven genes associated with diseases of the colon, particularly colon cancer, as identified by their coexpression with known colon cancer genes. The invention also relates to the use of these biomolecules in diagnosis, prognosis, prevention, treatment, and evaluation of therapies for diseases of the colon.

BACKGROUND ART

Colon cancer is the third leading cause of cancer deaths in the United States. Each year over 100,000 new cases are diagnosed, and 50,000 patients die from the disease. In large part this death rate is due to the inability to diagnose the disease at an early stage (Wanebo (1993) Colorectal Cancer, Mosby, St Louis MO). Although some of the genes that participate in or regulate the growth of colon cells are known, many other genes remain to be identified. Identification of new genes with significant levels of expression in cells of the diseased colon will provide new diagnostics, opportunities for earlier patient diagnosis, and targets for the development of therapeutic agents.

The present invention satisfies a need in the art by providing new compositions, seven genes associated with diseases of the colon identified by their coexpression patterns with genes expressed in colon cancer, that are useful for diagnosis, prognosis, treatment, prevention, and evaluation of therapies for diseases of the colon.

SUMMARY OF THE INVENTION

In one aspect, the invention provides for a substantially purified polynucleotide comprising a gene that is coexpressed with one or more known colon cancer genes in a plurality of biological samples. Preferably, known colon cancer genes are selected from the group consisting of carbonic anhydrase I, II, and IV (CA I, II, and IV), carcinoembryonic antigen family of proteins (cea), colorectal carcinoma tumor-associated antigen (CO-029), down-regulated in adenoma (dra), fatty-acid binding protein (fabp), galectin (galec), glutathione peroxidase (gpx2), guanylin (guan), cytokeratin 8 and 20 (ker 8 and 20), cadherin (cadher), and intestinal mucin (muc-2). Preferred embodiments include: (a) a polynucleotide sequence selected from SEQ ID NOs:1-7; (b) a polynucleotide sequence which encodes the polypeptide of SEQ ID NOs:8 or 9; (c) a polynucleotide sequence having at least 75% identity to the polynucleotide sequence of (a) or (b); (d) a polynucleotide sequence which is complementary to the polynucleotide sequence of (a), (b), or (c); (e) a polynucleotide sequence comprising at least 10, preferably at least 18, sequential nucleotides of the polynucleotide sequence of (a), (b), (c), or (d); or (f) a polynucleotide which hybridizes under stringent conditions to the polynucleotide of (a), (b), (c), (d) or (e). Furthermore, the invention provides an expression vector comprising any of the polynucleotides described above and host cells comprising the expression vector. Still further, the invention provides a method for treating or

preventing a disease or condition associated with the altered expression of a gene that is coexpressed with one or more known colon cancer genes comprising administering to a subject in need a polynucleotide described above in an amount effective for treating or preventing the disease.

In a second aspect, the invention provides a substantially purified polypeptide comprising the gene product of a gene that is coexpressed with one or more known colon cancer genes in a plurality of biological samples. The known colon cancer gene may be selected from the group consisting of carbonic anhydrase I, II, and IV, carcinoembryonic antigen family of proteins, colorectal carcinoma tumor-associated antigen, down-regulated in adenoma, fatty-acid binding protein , galectin, glutathione peroxidase, guanylin, cytokeratin 8 and 20, cadherin, and intestinal mucin. Preferred embodiments are (a) the polypeptide sequence of SEQ ID NOs:8 and 9; (b) a polypeptide sequence having at least 85% identity to the polypeptide sequence of (a); and (c) a polypeptide sequence comprising at least 6 sequential amino acids of the polypeptide sequence of (a) or (b). Additionally, the invention provides antibodies that bind specifically to any of the above described polypeptides and a method for treating or preventing a disease or condition associated with the altered expression of a gene that is coexpressed with one or more known colon cancer genes comprising administering to a subject in need such an antibody in an amount effective for treating or preventing the disease.

In another aspect, the invention provides a pharmaceutical composition comprising the polynucleotide of claim 2 or the polypeptide of claim 3 in conjunction with a suitable pharmaceutical carrier and a method for treating or preventing a disease or condition associated with the altered expression of a gene that is coexpressed with one or more known colon cancer genes comprising administering to a subject in need such a composition in an amount effective for treating or preventing the disease.

In a further aspect, the invention provides a method for diagnosing a disease or condition associated with the altered expression of a gene that is coexpressed with one or more known colon cancer genes, wherein each known colon cancer gene is selected from the group consisting of carbonic anhydrase I, II, and IV, carcinoembryonic antigen family of proteins, colorectal carcinoma tumor-associated antigen, down-regulated in adenoma, fatty-acid binding protein, galectin, glutathione peroxidase, guanylin, cytokeratin 8 and 20, cadherin, and intestinal mucin. The method comprises the steps of (a) providing a sample comprising one of more of the coexpressed genes; (b) hybridizing the polynucleotide of claim 2 to the coexpressed genes under conditions effective to form one or more hybridization complexes; (c) detecting the hybridization complexes; and (d) comparing the levels of the hybridization complexes with the level of hybridization complexes in a nondiseased sample, wherein altered levels of one or more of the hybridization complexes in a diseased sample compared with the level of hybridization complexes in a non-diseased sample correlates with the presence of the disease or

condition.

Additionally, the invention provides antibodies, antibody fragments, and immunoconjugates that exhibit specificity to any of the above described polypeptides and methods for treating or preventing diseases or conditions of the colon.

BRIEF DESCRIPTION OF THE SEQUENCE LISTING

The Sequence Listing provides exemplary colon cancer gene sequences including polynucleotide sequences, SEQ ID NOs:1-7, and the polypeptide sequences, SEQ ID NOs:8 and 9. Each sequence is identified by a sequence identification number (SEQ ID NO) and by the Incyte clone number with which the sequence was first identified.

DESCRIPTION OF THE INVENTION

It must be noted that as used herein and in the appended claims, the singular forms "a", "an", and "the" include the plural reference unless the context clearly dictates otherwise. Thus, for example, a reference to "a host cell" includes a plurality of such host cells, and a reference to "an antibody" is a reference to one or more antibodies and equivalents thereof known to those skilled in the art, and so forth.

DEFINITIONS

"NSEQ" refers generally to a polynucleotide sequence of the present invention, including SEQ ID NOs:1-7. "PSEQ" refers generally to a polypeptide sequence of the present invention, SEQ ID NOs:8 and 9.

A "fragment" refers to a nucleic acid sequence that is preferably at least 20 nucleic acids in length, more preferably 40 nucleic acids, and most preferably 60 nucleic acids in length, and encompasses, for example, fragments consisting of nucleic acids 1-50, 51-400, 401-4000, 4001-12,000, and the like, of SEQ ID NOs:1-7.

"Gene"refers to the partial or complete coding sequence of a gene and to its 5' or 3' untranslated regions. The gene may be in a sense or antisense (complementary) orientation.

"Colon cancer gene" refers to a gene whose expression pattern is similar to that of known colon cancer genes which are useful in the diagnosis, treatment, prognosis, or prevention of diseases of the colon, particularly colon cancer and other diseases associated with abnormal cell growth. "Known colon cancer gene" refers to a sequence which has been previously identified as useful in the diagnosis, treatment, prognosis, or prevention of diseases of the colon. Typically, this means that the known gene is expressed at higher levels (i.e., has more abundant transcripts) in diseased or cancerous colon tissue than in normal or non-diseased colon or any other tissue.

"Polynucleotide" refers to a nucleic acid molecule, nucleic acid sequence, oligonucleotide, nucleotide, or any fragment thereof. It may be DNA or RNA of genomic or synthetic origin,

double-stranded or single-stranded, and combined with carbohydrate, lipids, protein or other materials to perform a particular activity or form a useful composition. "Oligonucleotide" is substantially equivalent to the terms amplimer, primer, oligomer, element, and probe.

"Polypeptide" refers to an amino acid molecule, amino acid sequence, oligopeptide, peptide, or protein or portions thereof whether naturally occurring or synthetic.

A "portion" refers to peptide sequence which is preferably at least 5 to about 15 amino acids in length, most preferably at least 10 amino acids long, and which retains some biological or immunological activity of, for example, a portion of SEQ ID NOs:8 and 9.

"Sample" is used in its broadest sense. A sample containing nucleic acids may comprise a bodily fluid; an extract from a cell, chromosome, organelle, or membrane isolated from a cell; genomic DNA, RNA, or cDNA in solution or bound to a substrate; a cell; a tissue; a tissue print; and the like.

"Substantially purified" refers to a nucleic acid or an amino acid sequence that is removed from its natural environment and that is isolated or separated, and is at least about 60% free, preferably about 75% free, and most preferably about 90% free, from other components with which it is naturally present.

"Substrate" refers to any suitable rigid or semi-rigid support to which polynucleotides or polypeptides are bound and includes membranes, filters, chips, slides, wafers, fibers, magnetic or nonmagnetic beads, gels, capillaries or other tubing, plates, polymers, and microparticles with a variety of surface forms including wells, trenches, pins, channels, and pores.

A " variant" refers to a polynucleotide whose sequence diverges from SEQ ID NOs:1-7 or to a polypeptide who sequence diverges from SEQ ID NOs:8 and 9, respectively. Polynucleotide sequence divergence may result from mutational changes such as deletions, additions, and substitutions of one or more nucleotides; it may also be introduced to accommodate differences in codon usage. Each of these types of changes may occur alone, or in combination, one or more times in a given sequence. Polypeptide variants include sequences that possess at least one structural or functional characteristic of SEQ ID NOs:8 and 9.

THE INVENTION

The present invention encompasses a method for identifying biomolecules that are associated with a specific disease, regulatory pathway, subcellular compartment, cell type, tissue type, or species. In particular, the method identifies genes useful in diagnosis, prognosis, treatment, prevention, and evaluation of therapies for diseases of the colon including, but not limited, colon cancer, metastatic colon cancer, atrophic gastritis, cholecystitis, Crohns disease, irritable bowel syndrome, ulcerative colitis, and the like.

The method entails first identifying polynucleotides that are expressed in a plurality of cDNA libraries. The identified polynucleotides include genes of known or unknown function which are known

to be expressed in a specific disease process, subcellular compartment, cell type, tissue type, or species. The expression patterns of the genes with known function are compared with those of the genes with unknown function to determine whether a specified coexpression probability threshold is met. Through this comparison, a subset of the polynucleotides having a high coexpression probability with the known genes can be identified. The high coexpression probability correlates with a particular coexpression probability threshold which is preferably less than 0.001 and more preferably less than 0.00001.

The polynucleotides originate from cDNA libraries derived from a variety of sources including, but not limited to, eukaryotes such as human, mouse, rat, dog, monkey, plant, and yeast, and prokaryotes such as bacteria; and viruses. These polynucleotides can also be selected from a variety of sequence types including, but not limited to, expressed sequence tags (ESTs), assembled polynucleotide sequences, full length gene coding regions, promoters, introns, enhancers, 5' untranslated regions, and 3' untranslated regions. To have statistically significant analytical results, the polynucleotides need to be expressed in at least three cDNA libraries.

The cDNA libraries used in the coexpression analysis of the present invention can be obtained from adrenal gland, biliary tract, bladder, blood cells, blood vessels, bone marrow, brain, bronchus, cartilage, chromaffin system, colon, connective tissue, cultured cells, embryonic stem cells, endocrine glands, epithelium, esophagus, fetus, ganglia, heart, hypothalamus, immune system, intestine, islets of Langerhans, kidney, larynx, liver, lung, lymph, muscles, neurons, ovary, pancreas, penis, peripheral nervous system, phagocytes, pituitary, placenta, pleurus, prostate, salivary glands, seminal vesicles, skeleton, spleen, stomach, testis, thymus, tongue, ureter, uterus, and the like. The number of cDNA libraries selected can range from as few as 3 to greater than 10,000. Preferably, the number of the cDNA libraries is greater than 500.

In a preferred embodiment, genes are assembled to reflect related sequences, such as assembled sequence fragments derived from a single transcript. Assembly of the polynucleotide sequences can be performed using sequences of various types including, but not limited to, ESTs, extensions, or shotgun sequences. In a most preferred embodiment, the polynucleotide sequences are derived from human sequences that have been assembled using the algorithm disclosed in "System and Methods for Analyzing Biomolecular Sequences", USSN 09/276,534, filed March 25, 1999, incorporated herein by reference.

Experimentally, differential expression of the polynucleotides can be evaluated by methods including, but not limited to, differential display by spatial immobilization or by gel electrophoresis, genome mismatch scanning, representational difference analysis, and transcript imaging. Additionally, differential expression can be assessed by microarray technology. These methods may be used alone or in combination.

Known colon cancer genes can be selected based on the use of these genes as diagnostic or

prognostic markers or as therapeutic targets. Preferably, the known colon cancer genes include carbonic anhydrase I, II, and IV, carcinoembryonic antigen family of proteins, colorectal carcinoma tumor-associated antigen, down-regulated in adenoma, fatty-acid binding protein, galectin, glutathione peroxidase, guanylin, cytokeratin 8 and 20, cadherin, intestinal mucin, and the like.

The procedure for identifying novel genes that exhibit a statistically significant coexpression pattern with known colon cancer genes is as follows. First, the presence or absence of a gene in a cDNA library is defined: a gene is present in a cDNA library when at least one cDNA fragment corresponding to that gene is detected in a cDNA sample taken from the library, and a gene is absent from a library when no corresponding cDNA fragment is detected in the sample.

Second, the significance of gene coexpression is evaluated using a probability method to measure a due-to-chance probability of the coexpression. The probability method can be the Fisher exact test, the chi-squared test, or the kappa test. These tests and examples of their applications are well known in the art and can be found in standard statistics texts (Agresti (1990) Categorical Data Analysis, John Wiley & Sons, New York NY; Rice (1988) Mathematical Statistics and Data Analysis, Duxbury Press, Pacific Grove CA). A Bonferroni correction (Rice, supra, page 384) can also be applied in combination with one of the probability methods for correcting statistical results of one gene versus multiple other genes. In a preferred embodiment, the due-to-chance probability is measured by a Fisher exact test, and the threshold of the due-to-chance probability is set preferably to less than 0.001, more preferably to less than 0.00001.

To determine whether two genes, A and B, have similar coexpression patterns, occurrence data vectors can be generated as illustrated in Table 1. The presence of a gene occurring at least once in a library is indicated by a one, and its absence from the library, by a zero.

Table 1. Occurrence data for genes A and B

	Library 1	Library 2	Library 3	...	Library N
gene A	1	1	0	...	0
gene B	1	0	1	...	0

For a given pair of genes, the occurrence data in Table 1 can be summarized in a 2 x 2 contingency table.

Table 2. Contingency table for co-occurrences of genes A and B

	Gene A present	Gene A absent	Total
Gene B present	8	2	10
Gene B absent	2	18	20
Total	10	20	30

Table 2 presents co-occurrence data for gene A and gene B in a total of 30 libraries. Both gene A and gene B occur 10 times in the libraries. Table 2 summarizes and presents: 1) the number of times gene A and B are both present in a library, 2) the number of times gene A and B are both absent in a library, 3) the number of times gene A is present and gene B is absent, and 4) the number of times gene B is present and gene A is absent. The upper left entry is the number of times the two genes co-occur in a library, and the middle right entry is the number of times neither gene occurs in a library. The off diagonal entries are the number of times one gene occurs and the other does not. Both A and B are present eight times and absent 18 times. Gene A is present and gene B is absent two times; and gene B is present and gene A is absent two times. The probability ("p-value") that the above association occurs due to chance as calculated using a Fisher exact test is 0.0003. Associations are generally considered significant if a p-value is less than 0.01 (Agresti, supra; Rice, supra).

This method of estimating the probability for coexpression of two genes makes several assumptions. The method assumes that the libraries are independent and are identically sampled. However, in practical situations, the selected cDNA libraries are not entirely independent, because more than one library may be obtained from a single subject or tissue. Nor are they entirely identically sampled, because different numbers of cDNAs may be sequenced from each library. The number of cDNAs sequenced typically ranges from 5,000 to 10,000 cDNAs per library. In addition, because a Fisher exact coexpression probability is calculated for each gene versus 41,419 other assembled genes, a Bonferroni correction for multiple statistical tests is necessary.

Using the method of the present invention, we have identified seven novel genes that exhibit strong association, or coexpression, with known genes that are specific to colon cancer. These known colon cancer genes include carbonic anhydrase I, II, and IV, carcinoembryonic antigen family of proteins, colorectal carcinoma tumor-associated antigen, down-regulated in adenoma, fatty-acid binding protein, galectin, glutathione peroxidase, guanylin, cytokeratin 8 and 20, cadherin, and intestinal mucin. The results presented in Table 6 show that the expression of the seven novel genes have direct or indirect association with the expression of known colon cancer genes. Therefore, the novel genes can potentially be used in diagnosis, treatment, prognosis, or prevention of diseases of the colon or in the evaluation of therapies for diseases of the colon. Further, the gene products of the seven novel genes are either potential therapeutic proteins or targets of therapeutics against diseases of the colon.

Therefore, in one embodiment, the present invention encompasses a polynucleotide sequence comprising the sequence of SEQ ID NOs:1-7. These seven polynucleotides are shown by the method of the present invention to have strong coexpression association with known colon cancer genes and with each other. The invention also encompasses a variant of the polynucleotide sequence, its complement, or 18 consecutive nucleotides of a sequence provided in the above described sequences. Variant

polynucleotide sequences typically have at least about 75%, more preferably at least about 85%, and most preferably at least about 95% polynucleotide sequence identity to NSEQ.

NSEQ or the encoded PSEQ may be used to search against the GenBank primate (pri), rodent (rod), mammalian (mam), vertebrate (vrtp), and eukaryote (eukp) databases, SwissProt, BLOCKS (Bairoch et al. (1997) Nucleic Acids Res 25:217-221), PFAM, and other databases that contain previously identified and annotated motifs, sequences, and gene functions. Methods that search for primary sequence patterns with secondary structure gap penalties (Smith et al. (1992) Protein Engineering 5:35-51) as well as algorithms such as Basic Local Alignment Search Tool (BLAST; Altschul (1993) J Mol Evol 36:290-300; Altschul et al. (1990) J Mol Biol 215:403-410), BLOCKS (Henikoff and Henikoff (1991) Nucleic Acids Research 19:6565-6572), Hidden Markov Models (HMM; Eddy (1996) Cur Opin Str Biol 6:361-365; Sonnhammer et al. (1997) Proteins 28:405-420), and the like, can be used to manipulate and analyze nucleotide and amino acid sequences. These databases, algorithms and other methods are well known in the art and are described in Ausubel et al. (1997; Short Protocols in Molecular Biology, John Wiley & Sons, New York NY, unit 7.7) and in Meyers (1995; Molecular Biology and Biotechnology, Wiley VCH, New York NY, p 856-853).

Also encompassed by the invention are polynucleotide sequences that are capable of hybridizing to SEQ ID NOs:1-7, and fragments thereof under stringent conditions. Stringent conditions can be defined by salt concentration, temperature, and other chemicals and conditions well known in the art. Suitable conditions can be selected, for example, by varying the concentrations of salt in the prehybridization, hybridization, and wash solutions or by varying the hybridization and wash temperatures. With some substrates, the temperature can be decreased by adding formamide to the prehybridization and hybridization solutions.

Hybridization can be performed at low stringency, with buffers such as 5xSSC with 1% sodium dodecyl sulfate (SDS) at 60° C, which permits complex formation between two nucleic acid sequences that contain some mismatches. Subsequent washes are performed at higher stringency with buffers such as 0.2xSSC with 0.1% SDS at either 45° C (medium stringency) or 68° C (high stringency), to maintain hybridization of only those complexes that contain completely complementary sequences. Background signals can be reduced by the use of detergents such as SDS, Sarcosyl, or Triton X-100, and/or a blocking agent, such as salmon sperm DNA. Hybridization methods are described in detail in Ausubel (supra, units 2.8-2.11, 3.18-3.19 and 4-6-4.9) and Sambrook et al. (1989; Molecular Cloning, A Laboratory Manual, Cold Spring Harbor Press, Plainview NY)

NSEQ can be extended utilizing a partial nucleotide sequence and employing various PCR-based methods known in the art to detect upstream sequences such as promoters and other regulatory elements. (See, e.g., Dieffenbach and Dveksler (1995) PCR Primer, a Laboratory Manual, Cold Spring Harbor

Press, Plainview NY). Additionally, one may use an XL-PCR kit (PE Biosystems, Foster City CA), nested primers, and commercially available cDNA (Life Technologies, Rockville MD) or genomic libraries (Clontech, Palo Alto CA) to extend the sequence. For all PCR-based methods, primers may be designed using commercially available software, such as OLIGO 4.06 Primer analysis software (National Biosciences, Plymouth MN) or another appropriate program, to be about 18 to 30 nucleotides in length, to have a GC content of about 50%, and to form a hybridization complex at temperatures of about 68°C to 72°C.

In another aspect of the invention, NSEQ can be cloned in recombinant DNA molecules that direct the expression of PSEQ or structural or functional fragments thereof, in appropriate host cells. Due to the inherent degeneracy of the genetic code, other DNA sequences which encode substantially the same or a functionally equivalent amino acid sequence may be produced and used to express the polypeptide encoded by NSEQ. The nucleotide sequences of the present invention can be engineered using methods generally known in the art in order to alter the nucleotide sequences for a variety of purposes including, but not limited to, modification of the cloning, processing, and/or expression of the gene product. DNA shuffling by random fragmentation and PCR reassembly of gene fragments and synthetic oligonucleotides may be used to engineer the nucleotide sequences. For example, oligonucleotide-mediated site-directed mutagenesis may be used to introduce mutations that create new restriction sites, alter glycosylation patterns, change codon preference, produce splice variants, and so forth.

In order to express a biologically active protein, NSEQ, or derivatives thereof, may be inserted into an appropriate expression vector, i.e., a vector which contains the necessary elements for transcriptional and translational control of the inserted coding sequence in a particular host. These elements include regulatory sequences, such as enhancers, constitutive and inducible promoters, and 5' and 3' untranslated regions. Methods which are well known to those skilled in the art may be used to construct such expression vectors. These methods include in vitro recombinant DNA techniques, synthetic techniques, and in vivo genetic recombination. (See, e.g., Sambrook, supra; and Ausubel, supra).

A variety of expression vector/host cell systems may be utilized to express NSEQ. These include, but are not limited to, microorganisms such as bacteria transformed with recombinant bacteriophage, plasmid, or cosmid DNA expression vectors; yeast transformed with yeast expression vectors; insect cell systems infected with baculovirus vectors; plant cell systems transformed with viral or bacterial expression vectors; or animal cell systems. For long term production of recombinant proteins in mammalian systems, stable expression in cell lines is preferred. For example, NSEQ can be transformed into cell lines using expression vectors which may contain viral origins of replication and/or endogenous expression elements and a selectable or visible marker gene on the same or on a separate vector. The

invention is not to be limited by the vector or host cell employed.

In general, host cells that contain NSEQ and that express PSEQ may be identified by a variety of procedures known to those of skill in the art. These procedures include, but are not limited to, DNA-DNA or DNA-RNA hybridizations, PCR amplification, and protein bioassay or immunoassay techniques which include membrane, solution, or chip based technologies for the detection and/or quantification of nucleic acid or protein sequences. Immunological methods for detecting and measuring the expression of PSEQ using either specific polyclonal or monoclonal antibodies are known in the art. Examples of such techniques include enzyme-linked immunosorbent assays (ELISAs), radioimmunoassays (RIAs), and fluorescence activated cell sorting (FACS).

Host cells transformed with NSEQ may be cultured under conditions suitable for the expression and recovery of the protein from cell culture. The protein produced by a transgenic cell may be secreted or retained intracellularly depending on the sequence and/or the vector used. As will be understood by those of skill in the art, expression vectors containing NSEQ may be designed to contain signal sequences which direct secretion of the protein through a prokaryotic or eukaryotic cell membrane.

In addition, a host cell strain may be chosen for its ability to modulate expression of the inserted sequences or to process the expressed protein in the desired fashion. Such modifications of the polypeptide include, but are not limited to, acetylation, carboxylation, glycosylation, phosphorylation, lipidation, and acylation. Post-translational processing which cleaves a "prepro" form of the protein may also be used to specify protein targeting, folding, and/or activity. Different host cells which have specific cellular machinery and characteristic mechanisms for post-translational activities (e.g., CHO, HeLa, MDCK, HEK293, and WI38) are available from the American Type Culture Collection (ATCC, Manasas VA) and may be chosen to ensure the correct modification and processing of the expressed protein.

In another embodiment of the invention, natural, modified, or recombinant nucleic acid sequences are ligated to a heterologous sequence resulting in translation of a fusion protein containing heterologous protein moieties in any of the aforementioned host systems. Such heterologous protein moieties facilitate purification of fusion proteins using commercially available affinity matrices. Such moieties include, but are not limited to, glutathione S-transferase, maltose binding protein, thioredoxin, calmodulin binding peptide, 6-His, FLAG, *c-myc*, hemaglutinin, and monoclonal antibody epitopes.

In another embodiment, the nucleic acid sequences are synthesized, in whole or in part, using chemical or enzymatic methods well known in the art (Caruthers et al. (1980) Nucl Acids Symp Ser (7) 215-233; Ausubel, supra). For example, peptide synthesis can be performed using various solid-phase techniques (Roberge et al. (1995) Science 269:202-204), and machines such as the ABI 431A Peptide synthesizer (PE Biosystems) can be used to automate synthesis. If desired, the amino acid sequence may be altered during synthesis and/or combined with sequences from other proteins to produce a variant

protein.

In another embodiment, the invention entails a substantially purified polypeptide comprising the amino acid sequence of SEQ ID NOs:8 and 9 or fragments thereof.

DIAGNOSTICS and THERAPEUTICS

The polynucleotide sequences can be used in diagnosis, prognosis, treatment, prevention, and evaluation of therapies for diseases of the colon including, but not limited, colon cancer, metastatic colon cancer, atrophic gastritis, cholecystitis, Crohns disease, irritable bowel syndrome, ulcerative colitis, and the like.

In one preferred embodiment, the polynucleotide sequences are used for diagnostic purposes to determine the absence, presence, and excess expression of the protein. The polynucleotides may be at least 18 nucleotides long and consist of complementary RNA and DNA molecules, branched nucleic acids, and/or peptide nucleic acids (PNAs). In one alternative, the polynucleotides are used to detect and quantify gene expression in samples in which expression of NSEQ is correlated with disease. In another alternative, NSEQ can be used to detect genetic polymorphisms associated with a disease. These polymorphisms may be detected in the transcript cDNA.

The specificity of the probe is determined by whether it is made from a unique region, a regulatory region, or from a conserved motif. Both probe specificity and the stringency of diagnostic hybridization or amplification (maximal, high, intermediate, or low) will determine whether the probe identifies only naturally occurring, exactly complementary sequences, allelic variants, or related sequences. Probes designed to detect related sequences should preferably have at least 75% sequence identity to any of the nucleic acid sequences encoding PSEQ.

Methods for producing hybridization probes include the cloning of nucleic acid sequences into vectors for the production of mRNA probes. Such vectors are known in the art, are commercially available, and may be used to synthesize RNA probes in vitro by adding appropriate RNA polymerases and labeled nucleotides. Hybridization probes may incorporate nucleotides labeled by a variety of reporter groups including, but not limited to, radionuclides such as ^{32}P or ^{35}S, enzymatic labels such as alkaline phosphatase coupled to the probe via avidin/biotin coupling systems, fluorescent labels, and the like. The labeled polynucleotide sequences may be used in Southern or northern analysis, dot blot, or other membrane-based technologies; in PCR technologies; and in microarrays utilizing samples from subjects to detect altered PSEQ expression.

NSEQ can be labeled by standard methods and added to a sample from a subject under conditions suitable for the formation and detection of hybridization complexes. After incubation the sample is washed, and the signal associated with hybrid complex formation is quantitated and compared with a standard value. Standard values are derived from any control sample, typically one that is free of the

suspect disease. If the amount of signal in the subject sample is altered in comparison to the standard value, then the presence of altered levels of expression in the sample indicates the presence of the disease. Qualitative and quantitative methods for comparing the hybridization complexes formed in subject samples with previously established standards are well known in the art.

Such assays may also be used to evaluate the efficacy of a particular therapeutic treatment regimen in animal studies, in clinical trials, or to monitor the treatment of an individual subject. Once the presence of disease is established and a treatment protocol is initiated, hybridization or amplification assays can be repeated on a regular basis to determine if the level of expression in the subject begins to approximate that which is observed in a healthy subject. The results obtained from successive assays may be used to show the efficacy of treatment over a period ranging from several days to many years.

The polynucleotides may be used for the diagnosis of a variety of diseases associated with the colon. These include, but are not limited to, colon cancer, metastatic colon cancer, atrophic gastritis, cholecystitis, Crohns disease, irritable bowel syndrome, ulcerative colitis, and the like.

The polynucleotides may also be used as targets in a microarray. The microarray can be used to monitor the expression patterns of large numbers of genes simultaneously and to identify splice variants, mutations, and polymorphisms. Information derived from analyses of the expression patterns may be used to determine gene function, to understand the genetic basis of a disease, to diagnose a disease, and to develop and monitor the activities of therapeutic agents used to treat a disease. Microarrays may also be used to detect genetic diversity, single nucleotide polymorphisms which may characterize a particular population, at the genome level.

In yet another alternative, polynucleotides may be used to generate hybridization probes useful in mapping the naturally occurring genomic sequence. Fluorescent in situ hybridization (FISH) may be correlated with other physical chromosome mapping techniques and genetic map data as described in Heinz-Ulrich et al. (In: Meyers, supra, pp 965-968).

In another embodiment, antibodies or antibody fragments comprising an antigen binding site that specifically binds PSEQ may be used for the diagnosis of diseases characterized by the over-or-under expression of PSEQ. A variety of protocols for measuring PSEQ, including ELISAs, RIAs, and FACS, are well known in the art and provide a basis for diagnosing altered or abnormal levels of expression. Standard values for PSEQ expression are established by combining samples taken from healthy subjects, preferably human, with antibody to PSEQ under conditions suitable for complex formation The amount of complex formation may be quantitated by various methods, preferably by photometric means. Quantities of PSEQ expressed in disease samples are compared with standard values. Deviation between standard and subject values establishes the parameters for diagnosing or monitoring disease. Alternatively, one may use competitive drug screening assays in which neutralizing antibodies capable of

binding PSEQ specifically compete with a test compound for binding the protein. Antibodies can be used to detect the presence of any peptide which shares one or more antigenic determinants with PSEQ. In one aspect, the anti-PSEQ antibodies of the present invention can be used for treatment or monitoring therapeutic treatment for diseases of the colon, particularly colon cancer.

In another aspect, the NSEQ, or its complement, may be used therapeutically for the purpose of expressing mRNA and protein, or conversely to block transcription or translation of the mRNA. Expression vectors may be constructed using elements from retroviruses, adenoviruses, herpes or vaccinia viruses, or bacterial plasmids, and the like. These vectors may be used for delivery of nucleotide sequences to a particular target organ, tissue, or cell population. Methods well known to those skilled in the art can be used to construct vectors to express nucleic acid sequences or their complements. (See, e.g., Maulik et al. (1997) Molecular Biotechnology, Therapeutic Applications and Strategies, Wiley-Liss, New York NY.) Alternatively, NSEQ, or its complement, may be used for somatic cell or stem cell gene therapy. Vectors may be introduced in vivo, in vitro, and ex vivo. For ex vivo therapy, vectors are introduced into stem cells taken from the subject, and the resulting transgenic cells are clonally propagated for autologous transplant back into that same subject. Delivery of NSEQ by transfection, liposome injections, or polycationic amino polymers may be achieved using methods which are well known in the art. (See, e.g., Goldman et al. (1997) Nature Biotechnology 15:462-466.) Additionally, endogenous NSEQ expression may be inactivated using homologous recombination methods which insert an inactive gene sequence into the coding region or other appropriate targeted region of NSEQ. (See, e.g. Thomas et al. (1987) Cell 51:503-512.)

Vectors containing NSEQ can be transformed into a cell or tissue to express a missing protein or to replace a nonfunctional protein. Similarly a vector constructed to express the complement of NSEQ can be transformed into a cell to downregulate the overexpression of PSEQ. Complementary or antisense sequences may consist of an oligonucleotide derived from the transcription initiation site; nucleotides between about positions -10 and +10 from the ATG are preferred. Similarly, inhibition can be achieved using triple helix base-pairing methodology. Triple helix pairing is useful because it causes inhibition of the ability of the double helix to open sufficiently for the binding of polymerases, transcription factors, or regulatory molecules. Recent therapeutic advances using triplex DNA have been described in the literature. (See, e.g., Gee et al. In: Huber and Carr (1994) Molecular and Immunologic Approaches, Futura Publishing, Mt. Kisco NY, pp 163-177.)

Ribozymes, enzymatic RNA molecules, may also be used to catalyze the cleavage of mRNA and decrease the levels of particular mRNAs, such as those comprising the polynucleotide sequences of the invention. (See, e.g., Rossi (1994) Current Biology 4:469-471.) Ribozymes may cleave mRNA at specific cleavage sites. Alternatively, ribozymes may cleave mRNAs at locations dictated by flanking

regions that form complementary base pairs with the target mRNA. The construction and production of ribozymes is well known in the art and is described in Meyers (supra).

RNA molecules may be modified to increase intracellular stability and half-life. Possible modifications include, but are not limited to, the addition of flanking sequences at the 5' and/or 3' ends of the molecule, or the use of phosphorothioate or 2' O-methyl rather than phosphodiester linkages within the backbone of the molecule. Alternatively, nontraditional bases such as inosine, queosine, and wybutosine, as well as acetyl-, methyl-, thio-, and similarly modified forms of adenine, cytidine, guanine, thymine, and uridine which are not as easily recognized by endogenous endonucleases, may be included.

Further, an antagonist, or an antibody that binds specifically to PSEQ may be administered to a subject to treat or prevent a disease associated with colon cancer. The antagonist, antibody, or fragment may be used directly to inhibit the activity of the protein or indirectly to deliver a therapeutic agent to cells or tissues which express the PSEQ. An immunoconjugate comprising a PSEQ binding site of the antibody or the antagonist and a therapeutic agent may be administered to a subject in need to treat or prevent disease. The therapeutic agent may be a cytotoxic agent selected from a group including, but not limited to, abrin, ricin, doxorubicin, daunorubicin, taxol, ethidium bromide, mitomycin, etoposide, tenoposide, vincristine, vinblastine, colchicine, dihydroxy anthracin dione, actinomycin D, diphteria toxin, Pseudomonas exotoxin A and 40, radioisotopes, and glucocorticoid.

Antibodies to PSEQ may be generated using methods that are well known in the art. Such antibodies may include, but are not limited to, polyclonal, monoclonal, chimeric, and single chain antibodies, Fab fragments, and fragments produced by a Fab expression library. Neutralizing antibodies, such as those which inhibit dimer formation, are especially preferred for therapeutic use. Monoclonal antibodies to PSEQ may be prepared using any technique which provides for the production of antibody molecules by continuous cell lines in culture. These include, but are not limited to, the hybridoma, the human B-cell hybridoma, and the EBV-hybridoma techniques. In addition, techniques developed for the production of chimeric antibodies can be used. (See, e.g., Pound (1998) Immunochemical Protocols, Methods Mol Biol, Vol 80). Alternatively, techniques described for the production of single chain antibodies may be employed. Antibody fragments which contain specific binding sites for PSEQ may also be generated. Various immunoassays may be used to identify antibodies having the desired specificity. Numerous protocols for competitive binding or immunoradiometric assays using either polyclonal or monoclonal antibodies with established specificities are well known in the art.

Yet further, an agonist of PSEQ may be administered to a subject to treat or prevent a disease associated with decreased expression, longevity or activity of PSEQ.

An additional aspect of the invention relates to the administration of a pharmaceutical or sterile composition, in conjunction with a pharmaceutically acceptable carrier, for any of the therapeutic

applications discussed above. Such pharmaceutical compositions may consist of PSEQ or antibodies, mimetics, agonists, antagonists, or inhibitors of the polypeptide. The compositions may be administered alone or in combination with at least one other agent, such as a stabilizing compound, which may be administered in any sterile, biocompatible pharmaceutical carrier including, but not limited to, saline, buffered saline, dextrose, and water. The compositions may be administered to a subject alone or in combination with other agents, drugs, or hormones.

The pharmaceutical compositions utilized in this invention may be administered by any number of routes including, but not limited to, oral, intravenous, intramuscular, intra-arterial, intramedullary, intrathecal, intraventricular, transdermal, subcutaneous, intraperitoneal, intranasal, enteral, topical, sublingual, or rectal means.

In addition to the active ingredients, these pharmaceutical compositions may contain suitable pharmaceutically-acceptable carriers comprising excipients and auxiliaries which facilitate processing of the active compounds into preparations which can be used pharmaceutically. Further details on techniques for formulation and administration may be found in the latest edition of Remington's Pharmaceutical Sciences (Maack Publishing, Easton PA).

For any compound, the therapeutically effective dose can be estimated initially either in cell culture assays or in animal models such as mice, rats, rabbits, dogs, or pigs. An animal model may also be used to determine the appropriate concentration range and route of administration. Such information can then be used to determine useful doses and routes for administration in humans.

A therapeutically effective dose refers to that amount of active ingredient which ameliorates the symptoms or condition. Therapeutic efficacy and toxicity may be determined by standard pharmaceutical procedures in cell cultures or with experimental animals, such as by calculating and contrasting the ED_{50} (the dose therapeutically effective in 50% of the population) and LD_{50} (the dose lethal to 50% of the population) statistics. Any of the therapeutic compositions described above may be applied to any subject in need of such therapy, including, but not limited to, mammals such as dogs, cats, cows, horses, rabbits, monkeys, and most preferably, humans.

EXAMPLES

It is to be understood that this invention is not limited to the particular devices, machines, materials and methods described. Although particular embodiments are described, equivalent embodiments may be used to practice the invention. The described embodiments are not intended to limit the scope of the invention which is limited only by the appended claims. The examples below are provided to illustrate the subject invention and are not included for the purpose of limiting the invention.

I cDNA Library Construction

The COLNTUT16 cDNA library, in which Incyte clone 2790708 was discovered, was

constructed from colon tumor tissue obtained from a 60 year-old Caucasian male during a left hemicolectomy. Pathology indicated an invasive grade 2 adenocarcinoma, a sessile mass located three cm from the distal margin. The tumor extended through the submucosa and superficially into the muscularis propria. The margins of resection were free of involvement. One of nine regional lymph nodes contained metastatic adenocarcinoma. The patient presented with blood in the stool and a change in bowel habits. Patient history included thrombophlebitis, inflammatory polyarthropathy, prostatic inflammatory disease, and depressive disorder. Previous surgeries included resection of the rectum, a vasectomy, and exploration of the spinal canal. Family history included a malignant colon neoplasm in a sibling. The COLNNOT08 cDNA library in which Incyte clone 1843578 was discovered is from the same patient.

The frozen tissue was homogenized and lysed in TRIZOL reagent (1 gm tissue/10 ml TRIZOL; Life Technologies), a monoplastic solution of phenol and guanidine isothiocyanate, using a Polytron homogenizer (PT-3000; Brinkmann Instruments, Westbury NY). After a brief incubation on ice, chloroform was added (1:5 v/v), and the lysate was centrifuged. The chloroform layer was removed to a fresh tube, and the RNA extracted with isopropanol, resuspended in DEPC-treated water, and treated with DNase for 25 min at 37°C. The RNA was re-extracted once with acid phenol-chloroform pH 4.7 and precipitated using 0.3M sodium acetate and 2.5 volumes ethanol. The mRNA was isolated with the OLIGOTEX kit (Qiagen, Valencia CA) and used to construct the cDNA library.

The mRNA was handled according to the recommended protocols in the SUPERSCRIPT plasmid system (Life Technologies). The cDNAs were fractionated on a SEPHAROSE CL4B column (Amersham Pharmacia Biotech, Piscataway NJ), and those cDNAs exceeding 400 bp were ligated into pINCY 1 plasmid (Incyte Pharmaceuticals, Palo Alto CA). The plasmid was subsequently transformed into DH5α competent cells (Life Technologies).

II Isolation and Sequencing of cDNA Clones

Plasmid DNA was released from the cells and purified using the REAL Prep 96 plasmid kit (Qiagen). This kit enabled the simultaneous purification of 96 samples in a 96-well block using multi-channel reagent dispensers. The recommended protocol was employed except for the following changes: 1) the bacteria were cultured in 1 ml of sterile Terrific Broth (Life Technologies) with carbenicillin at 25 mg/L and glycerol at 0.4%; 2) after inoculation, the cultures were incubated for 19 hours; at the end of incubation, the cells were lysed with 0.3 ml of lysis buffer; and 3) following isopropanol precipitation, the plasmid DNA pellet was resuspended in 0.1 ml of distilled water, after which samples were transferred to a 96-well block for storage at 4° C.

The cDNAs were prepared using a MICROLAB 2200 (Hamilton, Reno NV) in combination with DNA ENGINE thermal cycler (PTC200; MJ Research, Watertown MA). cDNAs were sequenced by the

method of Sanger et al. (1975, J. Mol. Biol. 94:441f) using ABI PRISM 377 DNA sequencing systems (PE Biosystems) or MEGABASE 1000 sequencing systems (Molecular Dynamics, Sunnyvale CA).

Most of the sequences disclosed herein were sequenced using standard ABI protocols and ABI kits (Cat. Nos. 79345, 79339, 79340, 79357, 79355; PE Biosystems). The solution volumes were used at 0.25x -1.0x concentrations. Some of the sequences disclosed herein were sequenced using solutions and dyes from Amersham Pharmacia Biotech.

III Selection, Assembly, and Characterization of Sequences

The sequences used for coexpression analysis were assembled from EST sequences, 5' and 3' longread sequences, and full length coding sequences. Selected assembled sequences were expressed in at least three cDNA libraries.

The assembly process is described as follows. EST sequence chromatograms were processed and verified. Quality scores were obtained using PHRED (Ewing et al. (1998) Genome Res 8:175-185; Ewing and Green (1998) Genome Res 8:186-194), and edited sequences were loaded into a relational database management system (RDBMS). The sequences were clustered using BLAST with a product score of 50. All clusters of two or more sequences created a bin, and each bin with its resident sequences represents one transcribed gene.

Assembly of the component sequences within each bin was performed using a modification of Phrap, a publicly available program for assembling DNA fragments (Green, University of Washington, Seattle WA). Bins that showed 82% identity from a local pair-wise alignment between any of the consensus sequences were merged.

Bins were annotated by screening the consensus sequence in each bin against public databases, such as GBpri and GenPept from NCBI. The annotation process involved a FASTn screen against the gbpri database in GenBank. Those hits with a percent identity of greater than or equal to 75% and an alignment length of greater than or equal to 100 base pairs were recorded as homolog hits. The residual unannotated sequences were screened by FASTx against GenPept. Those hits with an E value of less than or equal to 10^{-8} were recorded as homolog hits.

Sequences were then reclustered using BLASTn and Cross-Match, a program for rapid protein and nucleic acid sequence comparison and database search (Green, supra), sequentially. Any BLAST alignment between a sequence and a consensus sequence with a score greater than 150 was realigned using cross-match. The sequence was added to the bin whose consensus sequence gave the highest Smith-Waterman score (Smith et al. supra) amongst local alignments with at least 82% identity. Non-matching sequences were moved into new bins, and assembly processes were performed for the new bins.

IV Coexpression Analyses of Known Colon Cancer Genes

Fourteen known colon cancer genes were selected to identify novel genes that are closely

associated with diseases of the colon. These known genes were carbonic anhydrase I, II, and IV, carcinoembryonic antigen family of proteins, colorectal carcinoma tumor-associated antigen, down-regulated in adenoma, fatty-acid binding protein, galectin, glutathione peroxidase, guanylin, cytokeratin 8 and 20, cadherin, and intestinal mucin. The colon cancer genes which were examined in this analysis and brief descriptions of their functions are listed in Table 4.

TABLE 4

GENE	DESCRIPTION AND REFERENCES
CA I, II, and IV	Carbonic anhydrase I, II, and IV Isoenzymes in colorectal mucosa, differentially expressed in colon cancer (Mori et al. (1993) Gastroenterology 105:820-6)
CEA	Carcinoembryonic antigen family of proteins Cell adhesion glycoprotein, diagnostic marker for colon cancer, prognostic for survival from colon cancer (Carpelan-Holmstrom et al. (1996) Dis Colon Rectum 39:799-805; Harrison et al. (1997) J Am Coll Surg 185:55-59; Graham et al. (1998) Ann Surg 228:59-63)
CO-029	CO-029 colorectal carcinoma tumor-associated antigen Cell surface glycoprotein (Sela et al. (1989) Hybridoma 8:481-491; Szala et al. (1990) Proc Natl Acad Sci 87:6833-6837)
DRA	Down-regulated in adenoma (DRA) Anion transporter expressed predominantly in colon mucosa, expression decreased in colon tumors, marker for progression of colon tumor (Schweinfest et al. (1993) Proc Natl Acad Sci 90:4166-4170; Byeon et al. (1996) Oncogene 12:387-396; Antalis et al. (1998) Clin Cancer Res 4:1857-1863)
FABP	Fatty-acid binding protein Hydrophobic ligand-binding protein expressed in liver and intestines, differentially expressed in colon and other cancers (Davidson et al. (1993) Lab Invest 68:663-675; Khan (1994) Proc Natl Acad Sci 91:848-852; Gromova et al. (1998) Int J Oncol 13:379-383)
Galec	Galectin family (Alternate name: IgE-binding protein) Modulate cell adhesion, cell proliferation, and cell death, differentially expressed in colon cancer including the metastatic phase (Sanjuan et al. (1997) Gastroenterology 113:1906-15; Bresalier et al. (1998) Gastroenterology 115:287-296; Perillo et al. (1998) J Mol Med 76:402-412)
Gpx2	Glutathione peroxidase Anti-oxidant, differentially expressed in colon cancers (Jendryczko et al. (1993) Neoplasma 40:107-109; Bravard et al. (1994) Int J Cancer 59:843-7; Beno et al. (1995) Neoplasma 42:265-9)
Guan	Guanylin Regulates chloride transport in epithelial tissues such as colon and shows decreased expression in colorectal adenocorcinoma (Cohen et al. (1998) Lab Invest 78:101-108)
ker 8 and 20	Cytokeratin 8 and 20 Cytoskeleton filaments and serum markers for colon cancer including the metastatic phase (Funaki, et al. (1997) Life Sci 60:643-652; Nakamori et al. (1997) Dis Colon Rectum 40: S29-36)

Cadher	Cadherin family Cell adhesion proteins and differentiation markers which are differentially expressed in colon and other cancers (Breen et al. (1995) Ann Surg Oncol 2:378-385; Eckert et al. (1997) Anticancer Res 17:7-12; Kreft, et al. (1997) J Cell Biol 136:1109-1121; Efstathiou et al. (1998) Proc Natl Acad Sci 95:3122-3127)
MUC-2	Intestinal mucin Expression decreased in majority of colorectal carcinomas (Ho et al. (1996) Oncol Res 8: 53-61; Hanski et al. (1997) J Pathol 182:385-391; Hanski et al. (1997) Lab. Invest. 77:685-95)

From a total of 41,419 assembled gene sequences, we have identified seven novel genes that show strong association with 14 known colon cancer genes. Initially, the degree of association was measured by probability values using a cutoff p value less than 0.00001. The sequences were further examined to ensure that the genes that passed the probability test had strong association with known colon cancer genes. The process was reiterated so that the initial 41,419 genes were reduced to the final seven colon disease associated genes. Details of the expression patterns for the 14 known and seven novel colon disease genes are presented in Tables 5 and 6.

Table 5 Co-Expression of the 14 Known Colon Cancer Genes (-log *p*)

		1	2	3	4	5	6	7	8	9	10	11	12	13
Guan	1													
Cadher	2	7												
CA IV	3	6	3											
FABP	4	13	8	4										
Galec	5	10	13	7	17									
CO-029	6	7	11	3	13	23								
DRA	7	13	10	10	20	21	17							
MUC-2	8	13	5	8	18	18	12	15						
CA I	9	15	4	5	11	7	5	9	8					
CEA	10	10	13	4	18	24	20	18	15	8				
Gpx2	11	8	12	5	16	25	19	15	11	6	21			
CA II	12	6	5	4	8	11	4	12	6	7	7	7		
ker20	13	14	10	7	16	21	19	18	16	10	24	18	7	
ker8	14	4	5	3	8	17	12	9	7	3	12	17	3	8

Table 6 Co-Expression of Seven Novel Genes and 14 Known Colon Cancer Genes (-log *p*)

Clone	Guan	Cadh	CA	FAB	Galec	CO-	DRA	MUC-	CA I	CEA	Gpx2	CA	ker20	ker8
2790708	8	4	3	5	6	3	8	3	4	4	5	3	4	2
1961467	2	3	1	4	4	2	8	4	2	3	4	3	3	3
1580553	5	4	6	12	12	8	10	15	5	13	12	4	15	5
2296694	2	3	3	2	7	9	2	1	1	6	7	1	3	16
1843578	10	5	3	7	6	3	8	7	8	5	4	5	8	2
2516888	14	6	6	20	21	13	17	16	8	14	14	7	15	8
3235282	10	8	5	12	16	12	17	10	9	14	18	8	15	7

We examined genes that are coexpressed with the 14 known colon cancer genes, and identified

seven novel genes that are strongly coexpressed. Each of the seven novel genes is coexpressed with at least one of the 14 known genes with a p-value of less than 10e-05. The coexpression of the seven novel genes with the 14 known genes are shown in Table 6. The entries in Table 6 are the negative log of the p-value (-log *p*) for the coexpression of the two genes. The novel genes identified are listed in the table by their Incyte clone numbers, and the known genes, by their abbreviated names as shown in Example V. For convenience, all the genes in the table 5 are assigned an identifying number, 1 to 14.

V Novel Genes Associated with Colon Diseases

Using the co-expression analysis method, we have identified seven novel genes that exhibit strong association, or co-expression, with 14 known colon cancer genes.

Nucleic acids comprising the consensus sequences of SEQ ID NOs:1-7 of the present invention were first identified from Incyte Clones 1580553, 1843578, 1961467, 2296694, 2516888, 2790708, and 32335282, respectively, and assembled according to Example III. BLAST and other motif searches were performed for SEQ ID NOs:1-7 according to Example VII. SEQ ID NOs:1-7 were translated and sequence identity was sought via comparison to known sequences. SEQ ID NOs:8 and 9 of the present invention were encoded by the nucleic acids of SEQ ID Nos:6-8, respectively. SEQ ID Nos:8 and 9 were also analyzed using BLAST and other motif search tools as disclosed in Example VI. Analyses of the novel genes is as follows.

SEQ ID NO:1 (Incyte clone 1580553) is 219 nucleotides in length and has about 74% identity to the nucleic acid sequence of a mouse mucin glycoprotein (g2583092). SEQ ID NO:2 (Incyte clone 2296694) is 252 nucleotides in length and has no known homologs in any of the public databases described in this application. SEQ ID NO:3 (Incyte clone 2516888) is 285 nucleotides in length and has no known homologs in any of the public databases described in this application. SEQ ID NO:4 (Incyte clone 2790708) is 1010 nucleotides in length and about 56% identity to the nucleic acid sequence from nucleotide 107789 to nucleotide 108777 of human chromosome 9 (g2564750). SEQ ID NO:5 (Incyte clone 3235282) is 2616 nucleotides in length and has about 64% identity to the nucleic acid sequence encoding a mouse calcium sensitive chloride conductance protein (g3925280) and 70% identity to a partial cDNAs of a colon specific gene, CSG5, which is 878 nucleotides long. SEQ ID NO:6 (Incyte clone 1843578) is 795 nucleotides in length and has about 64% identity to a nucleic acid sequence encoding a mouse calcium sensitive chloride conductance protein (g3925280). SEQ ID NO:7 (Incyte clone 1961467) is 2225 nucleotides in length and has about 6% identity to human gene signature HUMGS07792. SEQ ID NO:8 has 115 amino acids which are encoded by SEQ ID NO:6 and has no known homologs in any of the public databases described in this application. Motif analysis of SEQ ID NO:8 shows a potential phosphorylation site at S83. SEQ ID NO:9 has 90 amino acids which are encoded by SEQ ID NO:7 and has no known homologs in any of the public databases described in this

application. Motif analysis of SEQ ID NO:9 shows five potential phosphorylation sites at T10, T6, T21, S66, and S86.

VI Homology Searching for Colon Disease Genes and Their Encoded Proteins

The polynucleotide sequences, SEQ ID NOs:1-7, and polypeptide sequences, SEQ ID NOs:8 and 9, were queried against databases derived from sources such as GenBank and SwissProt. These databases, which contain previously identified and annotated sequences, were searched for regions of similarity using BLAST (Altschul, supra). BLAST searched for matches and reported only those that satisfied the probability thresholds of 10^{-25} or less for nucleotide sequences and 10^{-8} or less for polypeptide sequences.

The polypeptide sequences were also analyzed for known motif patterns using MOTIFS, SPSCAN, BLIMPS, and HMM-based protocols. MOTIFS (Genetics Computer Group, Madison WI) searches polypeptide sequences for patterns that match those defined in the Prosite Dictionary of Protein Sites and Patterns (Bairoch, supra) and displays the patterns found and their corresponding literature abstracts. SPSCAN (Genetics Computer Group) searches for potential signal peptide sequences using a weighted matrix method (Nielsen et al. (1997) Prot Eng 10:1-6). Hits with a score of 5 or greater were considered. BLIMPS uses a weighted matrix analysis algorithm to search for sequence similarity between the polypeptide sequences and those contained in BLOCKS, a database consisting of short amino acid segments, or blocks of 3-60 amino acids in length, compiled from the PROSITE database (Henikoff, supra; Bairoch, supra), and those in PRINTS, a protein fingerprint database based on non-redundant sequences obtained from sources such as SwissProt, GenBank, PIR, and NRL-3D (Attwood et al. (1997) J. Chem Inf Comput Sci 37:417-424). For the purposes of the present invention, the BLIMPS searches reported matches with a cutoff score of 1000 or greater and a cutoff probability value of 1.0×10^{-3}. HMM-based protocols were based on a probabilistic approach and searched for consensus primary structures of gene families in the protein sequences (Eddy, supra; Sonnhammer, supra). More than 500 known protein families with cutoff scores ranging from 10 to 50 bits were selected for use in this invention.

VII Labeling of Probes and Hybridization Analyses

Blotting

Polynucleotide sequences are isolated from a biological source and applied to a solid matrix (a blot) suitable for standard nucleic acid hybridization protocols by one of the following methods. A mixture of target nucleic acids is fractionated by electrophoresis through an 0.7% agarose gel in 1x TAE [40 mM Tris acetate, 2 mM ethylenediamine tetraacetic acid (EDTA)] running buffer and transferred to a nylon membrane by capillary transfer using 20x saline sodium citrate (SSC). Alternatively, the target nucleic acids are individually ligated to a vector and inserted into bacterial host cells to form a library.

Target nucleic acids are arranged on a blot by one of the following methods. In the first method, bacterial cells containing individual clones are robotically picked and arranged on a nylon membrane. The membrane is placed on bacterial growth medium, LB agar containing carbenicillin, and incubated at 37°C for 16 hours. Bacterial colonies are denatured, neutralized, and digested with proteinase K. Nylon membranes are exposed to UV irradiation in a STRATALINKER UV-crosslinker (Stratagene, La Jolla CA) to cross-link DNA to the membrane.

In the second method, target nucleic acids are amplified from bacterial vectors by thirty cycles of PCR using primers complementary to vector sequences flanking the insert. Amplified target nucleic acids are purified using SEPHACRYL-400 (Amersham Pharmacia Biotech). Purified target nucleic acids are robotically arrayed onto a glass microscope slide. The slide was previously coated with 0.05% aminopropyl silane (Sigma-Aldrich, St Louis MO) and cured at 110°C. The arrayed glass slide (microarray) is exposed to UV irradiation in a STRATALINKER UV-crosslinker (Stratagene).

Probe Preparation

cDNA probe sequences are made from mRNA templates. Five micrograms of mRNA is mixed with 1 μg random primer (Life Technologies), incubated at 70°C for 10 minutes, and lyophilized. The lyophilized sample is resuspended in 50 μl of 1x first strand buffer (cDNA Synthesis system; Life Technologies) containing a dNTP mix, [α-^{32}P]dCTP, dithiothreitol, and MMLV reverse transcriptase (Stratagene), and incubated at 42°C for 1-2 hours. After incubation, the probe is diluted with 42 μl dH_2O, heated to 95°C for 3 minutes, and cooled on ice. mRNA in the probe is removed by alkaline degradation. The probe is neutralized, and degraded mRNA and unincorporated nucleotides are removed using a PROBEQUANT G-50 MicroColumn (Amersham Pharmacia Biotech). Probes can be labeled with fluorescent markers, Cy3-dCTP or Cy5-dCTP (Amersham Pharmacia Biotech), in place of the radionuclide, [^{32}P]dCTP.

Hybridization

Hybridization is carried out at 65°C in a hybridization buffer containing 0.5 M sodium phosphate (pH 7.2), 7% SDS, and 1 mM EDTA. After the blot is incubated in hybridization buffer at 65°C for at least 2 hours, the buffer is replaced with 10 ml of fresh buffer containing the probe sequences. After incubation at 65°C for 18 hours, the hybridization buffer is removed, and the blot is washed sequentially under increasingly stringent conditions, up to 40 mM sodium phosphate, 1% SDS, 1 mM EDTA at 65°C. To detect signal produced by a radiolabeled probe hybridized on a membrane, the blot is exposed to a PHOSPHORIMAGER cassette (Molecular Dynamics), and the image is analyzed using IMAGEQUANT data analysis software (Molecular Dynamics). To detect signals produced by a fluorescent probe hybridized on a microarray, the blot is examined by confocal laser microscopy, and images are collected and analyzed using GEMTOOLS gene expression analysis software (Incyte Pharmaceuticals).

VIII Production of Specific Antibodies

SEQ ID NOs: 8-9, or portions thereof, substantially purified using polyacrylamide gel electrophoresis or other purification techniques, is used to immunize rabbits and to produce antibodies using standard protocols as described in Pound (supra).

Alternatively, the amino acid sequence is analyzed using LASERGENE software (DNASTAR, Madison WI) to determine regions of high immunogenicity, and a corresponding oligopeptide is synthesized and used to raise antibodies by means known to those of skill in the art. Methods for selection of appropriate epitopes, such as those near the C-terminus or in hydrophilic regions are well described in the art. Typically, oligopeptides 15 residues in length are synthesized using an ABI 431A Peptide synthesizer (PE Biosystems) using Fmoc-chemistry and coupled to keyhole limpet hemocyanin (KLH, Sigma-Aldrich) by reaction with N-maleimidobenzoyl-N-hydroxysuccinimide ester (Ausubel, supra) to increase immunogenicity. Rabbits are immunized with the oligopeptide-KLH complex in complete Freund's adjuvant. Resulting antisera are tested for antipeptide activity by, for example, binding the peptide to plastic, blocking with 1% BSA, reacting with rabbit antisera, washing, and reacting with radio-iodinated goat anti-rabbit IgG.

What is claimed is:

1. A substantially purified polynucleotide comprising a gene that is coexpressed with one or more known colon cancer genes in a plurality of biological samples, wherein each known colon cancer gene is selected from the group consisting of carbonic anhydrase I, II, and IV (CA I, II, and IV), carcinoembryonic antigen family of proteins (cea), colorectal carcinoma tumor-associated antigen (CO-029), down-regulated in adenoma (dra), fatty-acid binding protein (fabp), galectin (galec), glutathione peroxidase (gpx2), guanylin (guan), cytokeratin 8 and 20 (ker 8 and 20), cadherin (cadher), and intestinal mucin (muc-2).

2. RECONSTITUTE

(a) a polynucleotide sequence selected from the group consisting of SEQ ID NOs:1-7;

(b) a polynucleotide encoding a polypeptide sequence selected from the group consisting of SEQ ID NOs:8 and 9;

(c) a polynucleotide sequence having at least 75% identity to the polynucleotide sequence of (a) or (b);

(d) a polynucleotide sequence which is complementary to the polynucleotide sequence of (a), (b) or (c);

(e) a polynucleotide sequence comprising at least 18 sequential nucleotides of the polynucleotide sequence of (a), (b), (c), or (d); and

(f) a polynucleotide which hybridizes under stringent conditions to the polynucleotide of (a), (b), (c), (d), or (e).

3. A substantially purified polypeptide comprising the gene product of a gene that is coexpressed with one or more known colon cancer genes in a plurality of biological samples, wherein each known colon cancer gene is selected from the group consisting of carbonic anhydrase I, II, and IV (CA I, II, and IV), carcinoembryonic antigen family of proteins (cea), colorectal carcinoma tumor-associated antigen (CO-029), down-regulated in adenoma (dra), fatty-acid binding protein (fabp), galectin (galec), glutathione peroxidase (gpx2), guanylin (guan), cytokeratin 8 and 20 (ker 8 and 20), cadherin (cadher), and intestinal mucin (muc-2).

4. The polypeptide of claim 3, comprising a polypeptide sequence selected from the group consisting of:

(a) the polypeptide having the amino acid sequence selected from the group consisting of SEQ ID NOs:8 and 9;

(b) a polypeptide sequence having at least 85% identity to the polypeptide sequence of (a); and

(c) a polypeptide sequence comprising at least 6 sequential amino acids of the polypeptide sequence of (a) or (b).

5. An expression vector comprising the polynucleotide of claim 2.

6. A host cell comprising the expression vector of claim 5.

7. A pharmaceutical composition comprising the polynucleotide of claim 2 in conjunction with a suitable pharmaceutical carrier.

8. A pharmaceutical composition comprising the polypeptide of claim 3 in conjunction with a suitable pharmaceutical carrier.

9. An antibody or antibody fragment comprising an antigen binding site, wherein the antigen binding site specifically binds to the polypeptide of claim 4.

10. An immunoconjugate comprising the antigen binding site of the antibody or antibody fragment of claim 9 joined to a therapeutic agent.

11. A method for diagnosing a disease or condition associated with the altered expression of a gene that is coexpressed with one or more known colon cancer genes, wherein each known colon cancer gene is selected from the group consisting of carbonic anhydrase I, II, and IV (CA I, II, and IV), carcinoembryonic antigen family of proteins (cea), colorectal carcinoma tumor-associated antigen (CO-029), down-regulated in adenoma (dra), fatty-acid binding protein (fabp), galectin (galec), glutathione peroxidase (gpx2), guanylin (guan), cytokeratin 8 and 20 (ker 8 and 20), cadherin (cadher), and intestinal mucin (muc-2), the method comprising the steps of:

(a) providing a biological sample;

(b) hybridizing a polynucleotide of claim 2 to the biological sample under conditions effective to form one or more hybridization complexes;

(c) detecting the hybridization complexes; and

(d) comparing the levels of the hybridization complexes with the level of hybridization complexes in a non-diseased sample, wherein the altered level of hybridization complexes compared with the level of hybridization complexes of a nondiseased sample correlates with the presence of the disease or condition.

12. A method for treating or preventing a disease associated with the altered expression of a gene that is coexpressed with one or more known colon cancer genes in a subject in need, wherein each known colon cancer gene is selected from the group consisting of carbonic anhydrase I, II, and IV (CA I, II, and IV), carcinoembryonic antigen family of proteins (cea), colorectal carcinoma tumor-associated antigen (CO-029), down-regulated in adenoma (dra), fatty-acid binding protein (fabp), galectin (galec), glutathione peroxidase (gpx2), guanylin (guan), cytokeratin 8 and 20 (ker 8 and 20), cadherin (cadher), and intestinal mucin (muc-2), the method comprising the step of administering to the subject in need the pharmaceutical composition of claim 7 in an amount effective for treating or preventing the disease.

13. A method for treating or preventing a disease associated with the altered expression of a gene

that is coexpressed with one or more known colon cancer genes in a subject in need, wherein each known colon cancer gene is selected from the group consisting of carbonic anhydrase I, II, and IV (CA I, II, and IV), carcinoembryonic antigen family of proteins (cea), colorectal carcinoma tumor-associated antigen (CO-029), down-regulated in adenoma (dra), fatty-acid binding protein (fabp), galectin (galec), glutathione peroxidase (gpx2), guanylin (guan), cytokeratin 8 and 20 (ker 8 and 20), cadherin (cadher), and intestinal mucin (muc-2), the method comprising the step of administering to the subject in need the pharmaceutical composition of claim 8 in an amount effective for treating or preventing the disease.

14. A method for treating or preventing a disease associated with the altered expression of a gene that is coexpressed with one or more known colon cancer genes in a subject in need, wherein each known colon cancer gene is selected from the group consisting of carbonic anhydrase I, II, and IV (CA I, II, and IV), carcinoembryonic antigen family of proteins (cea), colorectal carcinoma tumor-associated antigen (CO-029), down-regulated in adenoma (dra), fatty-acid binding protein (fabp), galectin (galec), glutathione peroxidase (gpx2), guanylin (guan), cytokeratin 8 and 20 (ker 8 and 20), cadherin (cadher), and intestinal mucin (muc-2), the method comprising the step of administering to the subject in need the antibody or the antibody fragment of claim 9 in an amount effective for treating or preventing the disease.

15. A method for treating or preventing a disease associated with the altered expression of a gene that is coexpressed with one or more known colon cancer genes in a subject in need, wherein each known colon cancer gene is selected from the group consisting of carbonic anhydrase I, II, and IV (CA I, II, and IV), carcinoembryonic antigen family of proteins (cea), colorectal carcinoma tumor-associated antigen (CO-029), down-regulated in adenoma (dra), fatty-acid binding protein (fabp), galectin (galec), glutathione peroxidase (gpx2), guanylin (guan), cytokeratin 8 and 20 (ker 8 and 20), cadherin (cadher), and intestinal mucin (muc-2), the method comprising the step of administering to the subject in need the immunoconjugate of claim 10 in an amount effective for treating or preventing the disease.

16. A method for treating or preventing a disease associated with the altered expression of a gene that is coexpressed with one or more known colon cancer genes in a subject in need, wherein each known colon cancer gene is selected from the group consisting of carbonic anhydrase I, II, and IV (CA I, II, and IV), carcinoembryonic antigen family of proteins (cea), colorectal carcinoma tumor-associated antigen (CO-029), down-regulated in adenoma (dra), fatty-acid binding protein (fabp), galectin (galec), glutathione peroxidase (gpx2), guanylin (guan), cytokeratin 8 and 20 (ker 8 and 20), cadherin (cadher), and intestinal mucin (muc-2), the method comprising the step of administering to the subject in need the polynucleotide sequence of claim 2 in an amount effective for treating or preventing the disease.

SEQUENCE LISTING

```
<110> INCYTE PHARMACEUTICALS, INC.
      Walker, Michael, G.
      Volkmuth, Wayne
      Klingler, Tod, M.
      Lal, Preeti

<120> GENES ASSOCIATED WITH DISEASES OF THE COLON

<130> PB-0007 PCT

<140> To be assigned
<141> Herewith

<150> 09/255,381
<151> 1999-02-22

<160> 9

<170> PERL Program

<210> 1
<211> 219
<212> DNA
<213> Homo sapiens

<220>
<221> misc-feature
<223> Incyte ID No.: 1580553CB1

<400> 1
caccttctat atctctccag gctcaatgga aacaacatta gccagcacta ccacaacacc  60
aggcctcagt gcaaaatcta ccatccttta cagtagctcc agatcaccag accaaacact 120
ctcacctgcc agcatgagaa gctccagcat cagtggagaa cccaccagct tgtatagcca 180
agcagagtca acacacacaa cagcgttccc tgccagcac                        219


<210> 2
<211> 252
<212> DNA
<213> Homo sapiens

<220>
<221> unsure
<222> 201
<223> a or g or c or t, unknown, or other

<220>
<221> misc-feature
<223> Incyte ID No.: 2296694CB1

<400> 2
cttttcagaa ccccagatga gagccaatgt cagataaagt aagcatagca atgtagcagg  60
aactacaata gaagacattt tcactggaat tacaaagcag aattaaaatt atattgtaga 120
aggaaacacc aagaaaagaa tttccaggga aaatcctctt tgcaggtatt aattcttata 180
attttttgtc ttttggataa nctgtttact gcctcatctg aactgatccc aggtgaacgg 240
tttattgcct ag                                                     252


<210> 3
<211> 285
<212> DNA
<213> Homo sapiens
```

<220>
<221> misc-feature
<223> Incyte ID No.: 2516888CB1

<400> 3
gtggatgaca gggttggcca ccatggagca cctccaggct gacagagttg agacaagaac 60
ccatacctcc taactggcgc cactccaccc aggaggactc agccagccct tgagcacaca 120
gggacacact gctgaacctt atattgactt ccaatatgta tctttgctga gagaatgaat 180
gaaggaatga ttgtcagggg cactgccact gtgggggca tggccatcct ccaggtcact 240
gcggacttac ccctggccat ggcccagggc cctgctgtta ttatc 285

<210> 4
<211> 1010
<212> DNA
<213> Homo sapiens

<220>
<221> misc-feature
<223> Incyte ID No.: 2790708CB1

<400> 4
attttccttt actttttaaa taggttgttg cctcttatat atttattcta tgatgcaaat 60
gtcactatcc taattcctca gtttatgttt aacagcacac agtggcactt ctatgattca 120
aatacatttg ataacctttg aaatcaatca gaatactgca aaattaattt ttctaaaaca 180
atgcttttat cgttatttct cctgttgaat catcagtaca atttccaatt gaaaacactt 240
aaaataatct catattacaa tctttctcta acagaaccat gatgtaagga cagtgataac 300
aaatatctga caatgatatg attatttcct catccatgga aattttcctt aataaactaa 360
agggctattt tctaaaaagc caaagcattg cttacaagaa cttttcatca tgacatggat 420
agacactcag attcatacat tcaaagggaa gtgtcatgta ttccctttca atccacccta 480
ttctattgtg ttatcttcct aaattatttt ctatctacat tcttcattct ctttcccatt 540
gaccctatgt tctgtgtgat aaaaattgcg tcattggagg ctttttaagg ttaagtatta 600
tgccccattt caccattaat caacatacaa cccttctcca tattttgtaa ttcctttcat 660
atacagaaaa aaagatacta taatttcttc aaaatgcttg atattaatga tatatgggaa 720
aacaattatt ttgtgcagca atcttcagat aactgggaaa ggccggggaa aaagagagat 780
actggtggtt atcaatgacc catgtataaa ttgtttttat tatgtaagct gtcttcacaa 840
atgtcttctt atgtatgatc attagaactg ttttatatat atatgtaaaa tttccacatt 900
atcgagacat tactttcagc agtgaagtaa tcctttttta actgccactt aatgaattca 960
ataaaatata atttattgta ttttgctata ataaactatt gatgactatt 1010

<210> 5
<211> 2616
<212> DNA
<213> Homo sapiens

<220>
<221> misc-feature
<223> Incyte ID No.: 3235282CB1

<400> 5
aaaaatcgaa gcaacaaggt gttccgcagt atctctggta gaaatagagt ttataagtgt 60
caaggaggca gctgtcttag tagagcatgc agaattgatt ctacaacaaa actgtatgga 120
aaagattgtc aattctttcc tgataaagta caaacagaaa aagcatccat aatgtttatg 180
caaagtattg attctgttgt tgaattttgt aacgaaaaaa cccataatca agaagctcca 240
agcctacaaa acataaagtg caattttaga agtacatggg aggtgattag caattctgag 300
gattttaaaa acaccatacc catggtgaca ccacctcctc cacctgtctt ctcattgctg 360
aagatcagtc aaagaattgt gtgcttagtt cttgataagt ctggaagcat gggggtaag 420
gaccgcctaa atcgaatgaa tcaagcagca aaacatttcc tgctgcagac tgttgaaaat 480
ggatcctggg tggggatggt tcactttgat agtactgcca ctattgtaaa taagctaatc 540
caaataaaaa gcagtgatga aagaaacaca ctcatggcag gattacctac atatcctctg 600
ggaggaactt ccatctgctc tggaattaaa tatgcatttc aggtgattgg agagctacat 660
tcccaactcg atggatccga agtactgctg ctgactgatg gggaggataa cactgcaagt 720
tcttgtattg atgaagtgaa acaaagtggg gccattgttc attttattgc tttgggaaga 780
gctgctgatg aagcagtaat agagatgagc aagataacag gaggaagtca tttttatgtt 840
tcagatgaag ctcagaacaa tggcctcatt gatgcttttg ggctcttac atcaggaaat 900

actgatctct cccagaagtc ccttcagctc gaaagtaagg gattaacact gaatagtaat 960
gcctggatga acgacactgt cataattgat agtacagtgg gaaaggacac gttctttctc 1020
atcacatgga acagtctgcc tcccagtatt tctctctggg atcccagtgg aacaataatg 1080
gaaaatttca cagtggatgc aacttccaaa atggcctatc tcagtattcc aggaactgca 1140
aaggtgggca cttgggcata caatcttcaa gccaaagcga acccagaaac attaactatt 1200
acagtaactt ctcgagcagc aaattcttct gtgcctccaa tcacagtgaa tgctaaaatg 1260
aataaggacg taaacagttt ccccagccca atgattgttt acgcagaaat tctacaagga 1320
tatgtacctg ttcttggagc caatgtgact gctttcattg aatcacagaa tggacataca 1380
gaagttttgg aacttttgga taatggtgca ggcgctgatt ctttcaagaa tgatggagtc 1440
tactccaggt attttacagc atatacagaa aatggcagat atagcttaaa agttcgggct 1500
catggaggag caaacactgc caggctaaaa ttacggcctc cactgaatag agccgcgtac 1560
ataccaggct gggtagtgaa cggggaaatt gaagcaaacc cgccaagacc tgaaattgat 1620
gaggatactc agaccacctt ggaggatttc agccgaacag catccggagg tgcatttgtg 1680
gtatcacaag tcccaagcct tcccttgcct gaccaatacc caccaagtca aatcacagac 1740
cttgatgcca cagttcatga ggataagatt attcttacat ggacagcacc aggagataat 1800
tttgatgttg gaaaagttca acgttatatc ataagaataa gtgcaagtat tcttgatcta 1860
agagacagtt ttgatgatgc tcttcaagta aatactactg atctgtcacc aaaggaggcc 1920
aactccaagg aaagctttgc atttaaacca gaaaatatct cagaagaaaa tgcaacccac 1980
atatttattg ccattaaaag tatagataaa agcaatttga catcaaaagt atccaacatt 2040
gcacaagtaa ctttgtttat ccctcaagca aatcctgatg acattgatcc tacacctact 2100
cctactccta ctcctactcc tgataaaagt cataattctg gagttaatat ttctacgctg 2160
gtattgtctg tgattgggtc tgttgtaatt gttaacttta ttttaagtac caccatttga 2220
accttaacga agaaaaaaat cttcaagtag acctagaaga gagttttaaa aaacaaaaca 2280
atgtaagtaa aggatatttc tgaatcttaa aattcatccc atgtgtgatc ataaactcat 2340
aaaaataatt ttaagatgtc ggaaaaggat actttgatta aataaaaaca ctcatggata 2400
tgtaaaaact gtcaagatta aaatttaata gtttcattta tttgttattt tatttgtaag 2460
aaatagtgat gaacaaagat cctttttcat actgatacct ggttgtatat tatttgatgc 2520
aacagttttc tgaaatgata tttcaaattg catcaagaaa ttaaaatcat ctatctgagt 2580
agtcaaaata caagtaaagg agagcaaata aacatc 2616

<210> 6
<211> 795
<212> DNA
<213> Homo sapiens

<220>
<221> misc-feature
<223> Incyte ID No.: 1843578CB1

<400> 6
aggagaccca ggggtcccag agctgggctg gcgggaggcg taatccggcg gggtgagggt 60
tgatcgaaga gccccgcgcg cactgccgct cacagcccct tcccgagtgc agagcgggca 120
gagaagtcca ctgcttttaa ggccctgcac tgaaaatgca agctcaggcg ccggtggtcg 180
ttgtgaccca acctggagtc ggtcccggtc cggcccccca gaactccaac tggcagacag 240
gcatgtgtga ctgtttcagc gactgcggag tctgtctctg tggcacattt tgtttcccgt 300
gccttgggtg tcaagttgca gctgatatga atgaatgctg tctgtgtgga acaagcgtcg 360
caatgaggac tctctacagg acccgatatg gcatccctgg atctatttgt gatgactata 420
tggcaactct ttgctgtcct cattgtactc tttgccaaat caagagagat atcaacagaa 480
ggagagccat gcgtactttc taaaaactga tggtgaaaag ctcttaccga agcaacaaaa 540
ttcagcagac acctcttcag cttgagttct tcaccatctt ttgcaactga aatatgatgg 600
atatgcttaa gtacaactga tggcatgaaa aaaatcaaat ttttgattta ttataaatga 660
atgttgtccc tgaacttagc taaatggtgc aacttagttt ctccttgctt tcatattatc 720
gaatttcctg gcttataaac tttttaaatt acatttgaaa tataaaccaa atgaaatatt 780
aaaaaaaaaa aaaaa 795

<210> 7
<211> 2225
<212> DNA
<213> Homo sapiens

<220>
<221> misc-feature
<223> Incyte ID No.: 1961467CB1

<400> 7
gttcgggtcc tcggaccaca ctctggtttt ctatgctgtt ttggtgcaag tacaactgcc 60
gtagtcatgg ctttaggagc aataggattt taataaacag aacccatccc aaagccatga 120
ctacgacagt tgtacttgca ccaaaacagc atagaaaacc agagtgtggt gggaggaccc 180
gaagccggtt gggggaggat gtgagtaggg gcctggaggg tgcagggtca ttaatctgcg 240
gggagaacat tgtgctttag cccagggagg ggaggggtgg ggcaaatgca ccgaggtccc 300
cacttttcc tgctgccctc ggcaccctgg ggatgcaggc atctgggcac atctgcccct 360
tattgctgcc caccagcgtt aaacgccccc gatcccaaca ctagcaccac aggtggttcc 420
ggggcaggga gaggcaggaa tgggaaaatt gcttagagaa agattccact agaatccagt 480
gaattgtgct cagttctctt tacttcctac aaccgagtac atgggtcaca gggtggaggg 540
tgcaacagga catggaacat gcccctccgt gccccccaac acacacctgc acacaggatg 600
gtggtgtctg cagcatcaca ggtcatgcag ggcatgggga aggggaggtt cacacacaca 660
tagatgccca cagcgggtac cagacggaga acacccctga atatacatag ctgtacatgg 720
ggaaccccca ggtccccacc ccaaccctct cccctgtctt gctgtccccc gcaggggaac 780
tatattgctt tgagagagcc accccagggg ctgctctgcc aggcaccctc ccctcccacc 840
cacccccatt ttggcacatc tgcaagacac acagcagcga gagtaggcac cctcccttcc 900
caggcttctg tggcctggag ctggagaagg gggtaggaga cttcatcctc catcctcccc 960
taacccttcc caaacccctg ccaaacccac tcaagccaga acccaccccc acccccccaaa 1020
cacacataca aagctgagct atccaggaac acaagggaaa caaggagatt gtccagggtg 1080
ggagcggagg cagcgggga agaagactgg aagcagagac ctcccccctt gtgggggca 1140
gactggcaca acagctactt tagtgcaatt ggagagggtg cccagagtga gaggtggaga 1200
agggagggaa ggcggtcccc aacttccctg ggggcaaagt caggcttcca gattccccag 1260
ggaaagggcc tagcaggagt gggtgagggc caaggtggat cctctggtta cccgccaccc 1320
tctgccctcc caaatgcagt gacagtgtcc ccctcacacc taagtgggca acagcagcct 1380
tggagtcagt accttcaagt aattcaaaga gcagaccctc cccaccccag cttcacccca 1440
tctctgggat ttggtcgctt ctctaggggt tgggttggga ggagggagcc cccaaggcag 1500
acccttccct ctctacctcc cgattcccag accactgggc ttggtcctca aagattcctc 1560
acctccgccc ttgcccaacc tgggtcaagg ctgcagaagg ctggagccac cacaattaga 1620
ggggaagggg ctgctttgtt ccttatccct ccttcttaaa aggtagggtt caaactaggc 1680
gggatggggg cccatactgg tttgccccag gagtagggtt tctgggctag ggtctgtaag 1740
gctattttcc tttgcggtgg gaaggggagg taggggatga acactgggta tgggaagtgg 1800
gtgagaaatg gctgagaggg aaggaggaag gggcctcccc gctggagcag tcactggagt 1860
catttagaca aaaacactca tgtgcataag atacacagtg cgcaaactca gccctgccag 1920
cccggcccca atcccacctc tcaggactcc ttccaagacc ctggaggagg ttctggggat 1980
acagctgtag aaccgttcac tctggcccca tccaccccac ctccagcctc ttctcccctt 2040
ctaggtccag ggagtaagaa ggtgctcggg tggcagacca gtggtggaaa cagtattgag 2100
ttttcctttg gttacatatt gaaggcaaag gtgagctgga cttacagtca aaacggatag 2160
gggtgaggaa ggaagagggg ccatggctgg ggttggagag ggaggtaggc cctcgtcagc 2220
ccctc 2225

<210> 8
<211> 115
<212> PRT
<213> Homo sapiens

<220>
<221> misc-feature
<223> Incyte ID No.: 1843578CD1

<400> 8
Met Gln Ala Gln Ala Pro Val Val Val Val Thr Gln Pro Gly Val
1 5 10 15
Gly Pro Gly Pro Ala Pro Gln Asn Ser Asn Trp Gln Thr Gly Met
20 25 30
Cys Asp Cys Phe Ser Asp Cys Gly Val Cys Leu Cys Gly Thr Phe
35 40 45
Cys Phe Pro Cys Leu Gly Cys Gln Val Ala Ala Asp Met Asn Glu
50 55 60
Cys Cys Leu Cys Gly Thr Ser Val Ala Met Arg Thr Leu Tyr Arg
65 70 75
Thr Arg Tyr Gly Ile Pro Gly Ser Ile Cys Asp Asp Tyr Met Ala
80 85 90
Thr Leu Cys Cys Pro His Cys Thr Leu Cys Gln Ile Lys Arg Asp
95 100 105

```
Ile Asn Arg Arg Arg Ala Met Arg Thr Phe
                110                 115

<210> 9
<211> 90
<212> PRT
<213> Homo sapiens

<220>
<221> misc feature
<223> Incyte ID No.: 1961467CD1

<400> 9
Met Pro Thr Ala Gly Thr Arg Arg Arg Thr Pro Leu Asn Ile His
  1             5                   10                  15
Ser Cys Thr Trp Gly Thr Pro Arg Ser Pro Pro Gln Pro Ser Pro
                20                  25                  30
Leu Ser Cys Cys Pro Pro Gln Gly Asn Tyr Ile Ala Leu Arg Glu
                35                  40                  45
Pro Pro Gln Gly Leu Leu Cys Gln Ala Pro Ser Pro Pro Thr His
                50                  55                  60
Pro His Phe Gly Thr Ser Ala Arg His Thr Ala Ala Arg Val Gly
                65                  70                  75
Thr Leu Pro Ser Gln Ala Ser Val Ala Trp Ser Trp Arg Arg Gly
                80                  85                  90
```